STIC-ILL

QD1.A2

From: Clow, Lori
Sent: Monday, June 09, 2003 4:38 PM
To: STIC-ILL
Subject: request

I would like the following please:

Cawse, James N.
CS Combinatorial Chemistry Program, GE Corporate RD General Electric Company, Schenectady, NY, 12301, USA
SO Accounts of Chemical Research (2001), 34(3), 213-221

Materials A La Combi
Despite the challenges, a growing number of corporate and other labs see promise in studying materials combinatorially
SO Chemical & Engineering News, (15 May 2000) Vol. 78, No. 20, pp. 66.
CODEN: CENEAR, ISSN: 0009-2347.

TI Combinatorial and rapid screening approaches to homogeneous catalyst discovery and optimization
AU Crabtree, Robert H.
CS Yale Chemistry, New Haven, CT, 06520-8107, USA
SO Chemical Communications (Cambridge) (1999), (17), 1611-1616
CODEN: CHCOFS; ISSN: 1359-7345

Thanks

Lori A. Clow, Ph.D.
Patent Examiner, Art Unit 1631
Crystal Mall 1, Room 12-D-08
Mailbox 12-D-01
(703)-306-5439

Acc. Chem. Res. 2001, 34, 213–221

Experimental Strategies for Combinatorial and High-Throughput Materials Development

JAMES N. CAWSE
Combinatorial Chemistry Program, GE Corporate R&D, General Electric Company, P.O. Box 8, Schenectady, New York 12301

Received June 30, 2000

ABSTRACT
As high-throughput experimental techniques have become common in the area of materials research, entirely new types of experimental strategies have appeared. The kinds of problems, the desired outcomes, and the appropriate patterns are significantly different from those associated with conventional experimentation. Classical experimental design (design of experiments, DOE) strategies grew up in a period of slow, laborious, error-prone experimentation; a modern high-throughput laboratory can test more materials in a week than was previously done in a year. The goal of this Account is to identify and critically discuss some of the strategies that are being developed and used in this new, exciting area of research.

Introduction

Over the past 10 years, the new research technology called "combinatorial chemistry" or "high-throughput screening" has seen exponential growth. This technology—a set of techniques for creating a multiplicity of compounds and then testing them for activity—has been widely adopted in the pharmaceutical industry over the past few years. Virtually every major drug manufacturer is now using these techniques as the cornerstone of its research and development program. In the pharmaceutical industry, "libraries" of 1000 to 1 000 000 distinct compounds are routinely created and tested for biological activity. This is now practical because of the convergence of low-cost computer systems, reliable robotic systems, sophisticated molecular modeling, statistical experimental strategies, and database software tools.

In the last three to five years, this technology has expanded to materials design problems outside the drug field.[1] Major chemical companies have entered this arena, either by themselves or in concert with a company such as Symyx,[2] which specializes in new technologies for combinatorial materials discovery. Initial work has focused on development of robotic sample preparation, reactors, and sensors. Some of this equipment is becoming available commercially. With the use of this equipment, we have found that astonishing increases in the throughput of experimentation are possible (Figure 1).

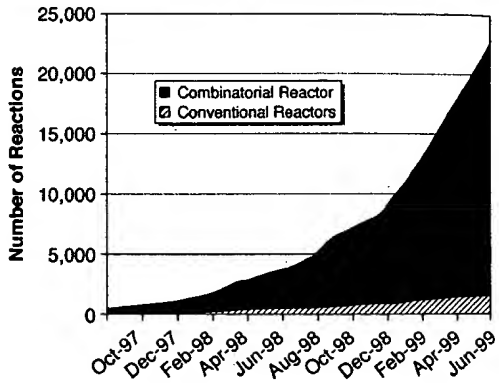

FIGURE 1. GE's experience with high-throughput screening of a catalyst system.

As our ability to generate large numbers of experiments has accelerated, we have become more conscious of the need to plan these experiments effectively. We find that the kinds of problems, the desired outcomes, and the appropriate strategies are significantly different from those associated with conventional experimentation. Classical experimental design strategies grew up in a period of slow, laborious, error-prone experimentation. The landmark designs developed by Fisher[3] were done in agricultural research where one experiment *per year* was the norm. Classic industrial design of experiments (DOE) studies[4] are usually attempts to determine the main effects and interactions of factors in a minimum number of experiments. These are now almost trivial; the emphasis is on the discovery of complex interactions by searching extensive chemical spaces.

Combinatorial Methods in the Scientific Landscape

The role of combinatorial methods in the general scientific landscape is one of scouting a wide array of possibilities for a low-probability "lead" to commercially interesting materials. This implies that the level of detailed scientific understanding of that area is relatively low; otherwise, more conventional experimentation would be more fruitful. Figure 2 gives a picture of the fit of combinatorial methods in the overall range of scientific strategies. As the level of scientific understanding of a problem increases, the quality of the equations and mathematical models we use to represent that understanding also increases. Consequently, the kinds of experiments we perform to generate data also change. At the lowest level of knowledge, where we only have a first insight into a potentially attractive chemical "universe", empirical strategies such as combinatorial methods are most attractive. As the system becomes better known, the number of potentially important factors and their ranges will decrease. More conventional strategies such as the widely

Dr. James N. Cawse graduated from Wesleyan University in 1967 and received his Ph.D. in chemistry from Stanford University in 1973. After six years with Union Carbide in South Charleston, WV, he joined GE Plastics in Pittsfield, MA, as a process development chemist. His career gradually shifted from chemistry to statistics, and in 1990 he led the applied statistics program at GE Plastics. When GE began the Six-Sigma quality program, he immediately became a Black Belt Instructor. In 1998 he moved to GE Corporate R&D in Schenectady, NY, as the data team leader for the combinatorial chemistry program. He has published 17 papers and holds 5 patents.

10.1021/ar000117s CCC: $20.00
Published on Web 01/03/2001

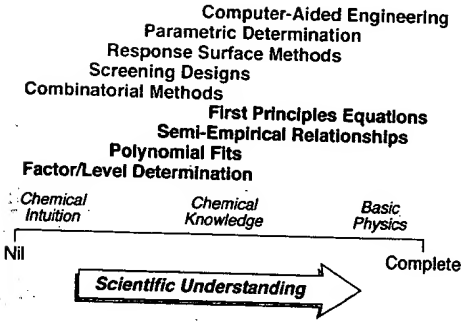

FIGURE 2. As scientific understanding increases, the mathematical models (black) used to describe a phenomenon and the experimental techniques (gray) used to sample it become more sophisticated.

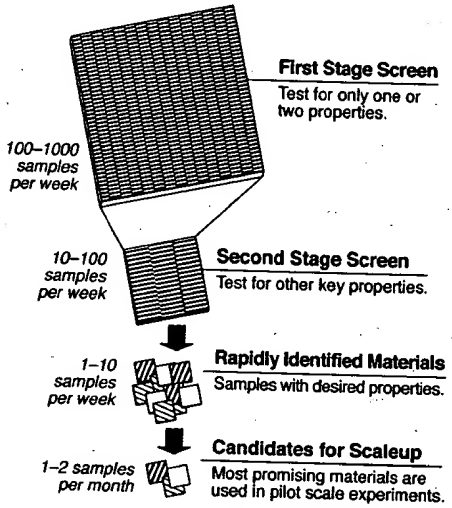

FIGURE 3. High-throughput methodologies require a highly structured approach to achieve the productivity improvements advertised. Multistage screening must be integrated with laboratory- and pilot-scale testing.

used factorial and response surface designs[5] will then become more appropriate.

All of these programs in combinatorial or high-throughput materials development use some form of a multiphase strategy (Figure 3). A first-stage screen may only test for one or two critical properties which are easily and quickly measured on a microscale. This may be followed up by a second screen, also on a microscale, to test for other key properties or optimize the settings of the process parameters. The best materials will be tested on a standard laboratory scale where such parameters as mass balance are more accurately determined. Finally, a very few materials will become candidates for scale-up in a pilot facility. All of these efforts occur in the larger context of learning about the overall chemical system. The information obtained from the combinatorial experiment is fed back to the design process in the form of appropriate descriptors of the experimental space. These descriptors can be used to structure or constrain the space so the experimental process converges more quickly. The resources for all of these phases must be in balance so there are no bottlenecks in the testing process. In addition, all of the steps in the screening process must be in proper balance. In a high-throughput process, you must "analyze in a day what you make in a day".[6]

The goals and strategies of combinatorial techniques applied to materials development are quite different from those of the pharmaceutical arena. Some of these differences are given in Table 1. The primary goal of pharmaceutical research is development of a single compound that is effective as a drug. The total number of druglike molecules is estimated to exceed 10^{64} possibilities. This leads to a focus of combinatorial drug strategy: "which small portion of all accessible compounds should be made to have the greatest chance of progressing the drug design project?"[7] The most common current strategy is one of *diversity*: selecting a subset of compounds which represent the "chemical space" under investigation. This, in turn, requires metrics that describe the chemical space; these are typically derived from properties which can easily be calculated from the structure of the compounds being studied.[8]

In materials development, the primary goal is discovery of systems that meet a number of physical, chemical, and structural requirements. These systems may be catalysts, polymers, phosphors, electronic materials, pigments, or coatings.[9] Such systems are likely to involve several molecular species and process variables. An industrially interesting materials development problem will typically have been the subject of years (or decades!) of conventional research; in that work, all the primary effects and simple interactions of the various parameters of the system will have been investigated. If something new is to be found in the system, it will be from the synergistic effects of three or more parameters working together (Figure 4).[10] The probability of finding such three-way or higher interactions is too low for them to be likely to be found by conventional means. Only high-throughput experimentation will be able to find them.

Even with high-throughput methodology, however, the combinatorial explosion of possibilities represents a daunting task. For example, in a relatively simple single-phase homogeneous catalyst system, the number of possible experiments quickly rises into the millions (Table 2). If we add the complications of multiple phases, as would occur in a heterogeneous catalyst, the possibilities grow even more numerous.

Table 1.

pharmaceutical	materials development
• focused on chemical synthesis as primary	• synthesis, mixtures, and process variables
• emphasis on diversity within known metrics	• emphasis on broad coverage and synergy
• experimental space metrics known	• experimental space metrics not known
• easy sample evaluation on nanogram level	• sample evaluation difficult and individual for each system
• challenge is designing diverse libraries from very large numbers ($\gg 10^6$) of molecules	• challenge is finding high order synergies of qualitative and mixture/process variables

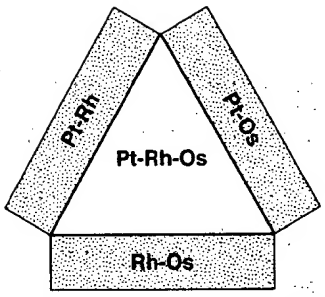

FIGURE 4. Highly active ternary catalyst bounded by low-activity binaries.

Table 2. Possible Numbers of Experiments in a Representative Situation

	type	levels
formulation factors		
primary catalyst	qualitative	1
inorganic cocatalyst	qualitative	20
amount of cocatalyst	quantitative	3
organic ligand	qualitative	20
amount of ligand	quantitative	3
active anion	qualitative	10
amount of anion	quantitative	3
process factors		
reaction time	quantitative	3
reaction temperature	quantitative	3
reaction pressure	quantitative	3
total number of potential runs		2 916 000

Within this complex area of research, I would suggest a few reasonable goals for the experimental strategist. We need

- strategies to address very large, multidimensional experimental spaces
- a taxonomy of the varieties of experimental spaces
- estimates of what is discoverable...and what is not
- decision rules for deciding when to *stop* studying a space
- predictive methods for generating fruitful experiments.

The focus in this work will be on the first three of these points.

Experimental Strategy in Combinatorial Organic Synthesis

Work in the pharmaceutical industries has led to considerable discussion of experimental strategy in this area.[11−19] These articles have mostly focused on what can broadly be called "diversity strategies",[18] in which

- structural descriptors are calculated for each compound in a potential library;
- similarity coefficients are calculated between compound pairs; and
- compounds are selected for libraries using cluster-based, dissimilarity-based, or partition-based methods.

These methodologies have often been compared against pure random screening.[15] The advantages and disadvantages of each method are still a subject of active debate.

The crucial advantage that combinatorial organic synthesis has over materials development is its focus on single compounds as targets. From these target molecules, descriptors can be calculated and used as metrics in a quantitative experimental space. This space is quite large, ranging from 19 dimensions[16] to thousands, but it is possible to generate rational diversity within that space using the methods mentioned above.

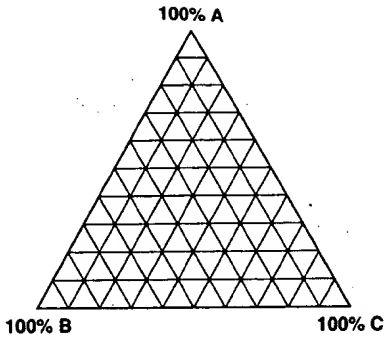

FIGURE 5. Ternary gradient in 10% steps.

Approaches to Experimental Strategy in Materials Development

High-Speed Array Strategies. The properties of functional materials such as phosphors, catalysts, and electronic components arise from complex interactions of their formulation and processing. The development of descriptors in these areas is in its infancy. For that reason, descriptor-based experimental strategies in these areas have tended to be limited. "There is no approach that will have the generality of the combinatorial methods currently used for synthesis and screening of biologically active molecules."[20] In fact, many "combinatorial" materials development programs are best characterized as array methods for rapid performance of conventional experiments. In the following section, I will discuss some of the more common approaches in the literature from an experimental strategy viewpoint.

1. Gradient Arrays. A common approach in solid-state materials studies is examination of a ternary (or higher) materials gradient.[21] This can be done by using continuous or point techniques. In "continuous composition spread",[22] a single film with a ternary composition spread was generated on a 63- × 66-mm substrate in one step, and the electronic properties were measured at ~4000 points. It found an excellent dielectric $Zr_{0.15}Sn_{0.3}Ti_{0.55}O_{2-\delta}$. The experimental design and strategy issues in this type of experiment are limited to the choice of experimental system and the fineness of the test gradient. This has been most developed in the study of electronic thin-film materials. It is particularly suited for identification of narrow phase regions with suitable properties. It is dependent on very fast, high-resolution methods of property determination.

A more common theme is a ternary gradient studied at regular intervals. Intervals such as 0−100% by 10% steps or 0−1% by 0.1% steps are convenient; these generate 66-point triangular arrays (Figure 5). For example:

- The Pt−Pd−In system for cyclohexane dehydrogenation catalysis was studied in the 0−1% range at 0.1% intervals.[23]

Table 3. Representative Metal Oxide Materials (Host:Activator)

host metal atoms	phosphors	superconductors	magnetoresistance
1	$Y_2O_3:Eu^{3+}$		Fe_2O_4:Pd
2	$Y_3Al_5O_{12}:Ce^{3+}$	La_2CuO_4	$La_{0.67}Ca_{0.33}MnO_3$
3	$BaMgAl_{10}O_{17}:Eu^{2+}$	$YBa_2Cu_3O_7$	

• The Rh–Pd–Pt system for CO oxidation was studied over the 0–100% range using 15 steps.[24]

The technique can be extended to more complex combinations. A quaternary phase diagram was studied in the Pt/Ru/Os/Ir system for methanol fuel cell catalysis.[10] Typically, the response is a visual signal (either directly or indirectly), and the analysis of the data has often been done by visual inspection.

In these designs, the overall shape is determined, but there are still important strategy decisions to be made:

• What grid density should be used? Typically these arrays are designed to locate a relatively small region of phase space in which a phase with advantageous properties is located. The grid density will determine the smallest phase space that can be observed. The tradeoff, of course, is that halving the distance between levels almost quadruples the number of samples.

• Should the grid have uniform spacing? In solid-state chemistry, it is common for a component to have its most important effect as a dopant at very small concentrations. Uniform spacing may oversample the center of the space while missing the potential dopant regions at the edge. A logarithmic spacing (e.g., 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100%) which places more points in the low-concentration region may be advantageous.

• How do we decide whether the quality of an array is good enough to make it usable? An advantage of gradient arrays (unlike quaternary mask arrays) is that there is a direct geometric concentration gradient, so that trends can be located visually or with curve-fitting techniques. Quality criteria must be set for decisions based on lack of trends or randomness.

• How will we detect a "hit" in the array? A "pick-the-winner" strategy is the simplest but also most likely to be fooled by noise in the data. For screening purposes, entire areas of high-response compositions can be selected and made into "focussed" arrays.

2. Quaternary Mask Arrays. These designs were developed to exploit the unique features of the inorganic chemistry of metal oxides. A large number of scientifically and commercially important materials such as phosphors, scintillators,[25] light-emitting diodes, and superconductors are composed of a metal oxide host lattice doped with small amounts of other metal atoms as activators (Table 3).

These structures can be summarized as $A_mB_nO_x$:dopant for two host atoms; $A_mB_nC_oO_x$:dopant for three, and so on. The metal atoms must be chosen for the A, B, and C positions with size and charge appropriate to the crystal structure being built. When thin layers of metal oxides are deposited onto a substrate and carefully annealed, the crystal structures form spontaneously.

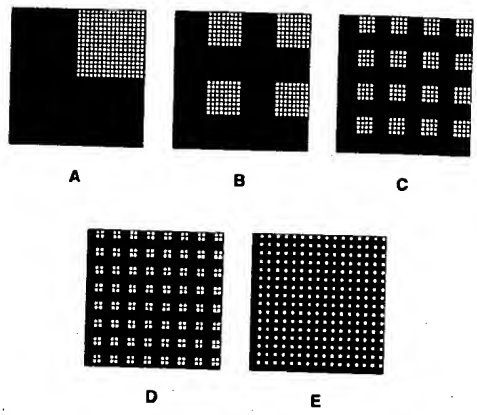

FIGURE 6. Six fractal quaternary masks used in deposition studies.

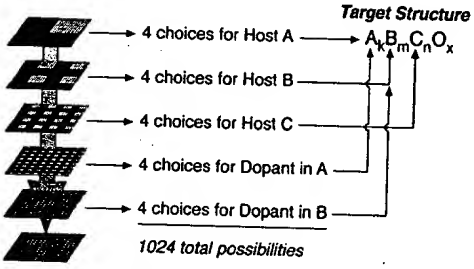

FIGURE 7. The quaternary masking system uses six fractal masks, each of which can be rotated 90° to allow four choices of material at each level.

Since there can be many choices for metals in each of the A, B, C, and dopant positions, the quaternary mask system[26] (Figure 6) was designed to enable free and flexible choices in each position. In its current state of development, 1024 distinct samples can be generated in just 24 sputtering operations using six masks. Each mask can be used four times by 90° rotations. This allows four choices for each host position and each dopant (Figure 7).

The quaternary mask system has important advantages over its predecessor, the binary mask system.[27] Binary masks do not allow efficient separation of metals by function, so a large fraction of the samples made have compositions which do not form the correct structure. In all these masking systems, an additional degree of compositional freedom can be added by gradually moving the mask or a shutter during the deposition procedure.[28]

3. High-Speed Versions of Conventional Experimental Designs. These designs are frequently used in the second stage of high-throughput screening, when a "hit" has been located. Since the cost of experimental points is relatively low, these can be quite high resolution designs such as full factorials, central composite designs,[5] and special cubic or cubic mixture designs.[29] The classic experimental design issue which frequently crops up in these experiments is *nesting*. It is generally quite easy to make an array of compositions in one of the standard designs; however, these arrays are usually subjected to physical treatments (heating, cooling, gas pressure, etc.) as units. The composition variables are therefore nested[30] within the physical treatment variables, and appropriate designs and analyses must be used.[31–33]

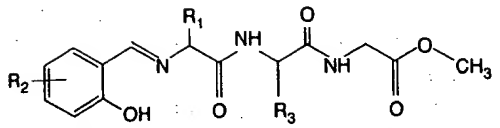

FIGURE 8. Scaffold for synthesis of enantioselective catalyst using representational catalyst strategy. R_1, R_2, and R_3 are variable substituents.

True Combinatorial Design Strategies. There are two possible interpretations of the term "combinatorial" in experimental situations. The one used in the pharmaceutical arena is in the sense of actual combinations of compounds in the course of experimentation. This is most obvious in the now classic "split-and-pool" technique, in which polymer-bound compounds are split into separate vessels where each is treated with a different reagent and then recombined into a common pool. Repetition of this process yields a mixture of m^n compounds, where m is the number of separate vessels at each step and n is the number of steps. This has also been used in some catalyst development programs where the catalyst is a single organic species—analogous to an enzyme.[34]

A second meaning, which I emphasize here, is the use of combinatorial mathematics to calculate and sample the possible combinations of parameters in a materials system. If, as noted above, the potential groundbreaking material innovations of the next generation will be found in high-level synergies, it makes sense to use appropriate mathematical tools to locate them.

The use of these mathematical combinatorial methods will be most prevalent in the early stages of a project, before the investigators have been able to develop appropriate descriptors. In such a situation, we only have a set of potentially important factors, each of which may have many levels. Several strategies have been used so far in searching for potential valuable synergies.

1. "Representational" Strategy.[35] In this approach, a molecular catalyst containing three variable substituents (Figure 8) was to be optimized. There were 20 possible substituents in each region, so the total number of possibilities was $20^3 = 8000$. Rather than test all 8000, the 20 possible variations in the first variable region were tested and the best selected. It was then fixed and the 20 possibilities for the second region were tested, followed by fixing the best and testing the third region. This is illustrated in Figure 9.

While this did find a substantially improved catalyst using only 60 of 8000 possible experiments, it is also a very limited strategy. It is entirely analogous to the "one-variable-at-a-time" strategies in conventional experimentation.[5] If there are any interactions between the three variable regions, they will not be found.

2. Index Library Strategy.[36] This method was used to find an optimal metal–ligand catalyst combination. Given 10 ligands and 10 metals, 100 combinations are possible. Instead, the 10 ligands were mixed together, and the mixture was tested, one at a time, with each of the metals. Similarly, the 10 metals were mixed together, and the mixture was tested with each of the ligands. The best

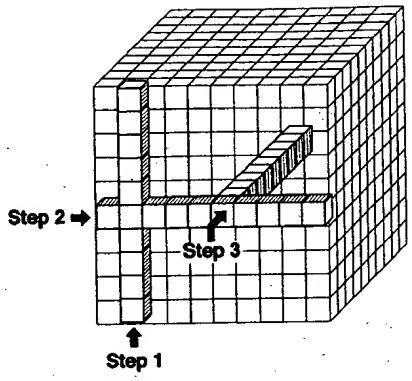

FIGURE 9. "Representational" catalyst search strategy.

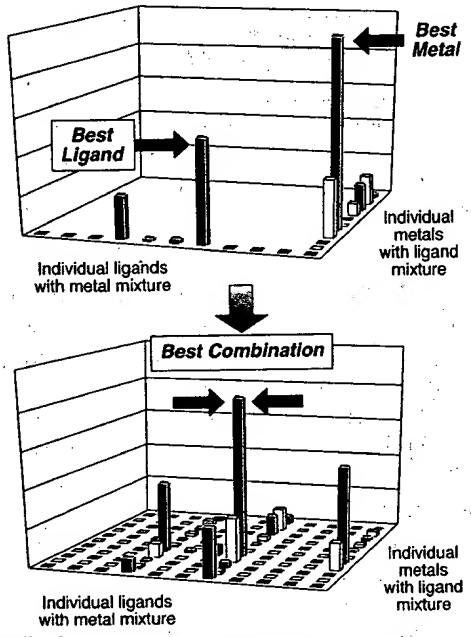

FIGURE 10. "Index library" catalyst search strategy.

results from the two sets of experiments then indicated the best metal–ligand pair (Figure 10).

This, too, is a very limited strategy. It can only be used with relatively small metal–ligand or similar systems. If there are too many of either one, the concentration of the active species will be diluted to the point where it will not appear above the noise. Cross reactions or competition in which multiple different ligands bond to a single metal may be possible. Finally, these kinds of systems will frequently contain catalyst poisons as well as catalysts. This will severely impact the usefulness of the method.

3. All Two-Way Combinations Strategy.[37] This method was used to find optimal catalyst systems in a situation where it had been found that combinations of metal cocatalysts were advantageous. Nineteen possible metals were identified, and all possible pairs of catalysts were tried (Figure 11). Eleven systems were identified as having possible synergy and were passed to secondary testing.

This strategy also has its limitations. Only two-way combinations were tested here. If three-way combinations were potentially interesting, the number of tests jumped to $(19 \times 18 \times 17)/(1 \times 2 \times 3) = 969$, which was too many for the budget. Also, the metals were only tested at a single

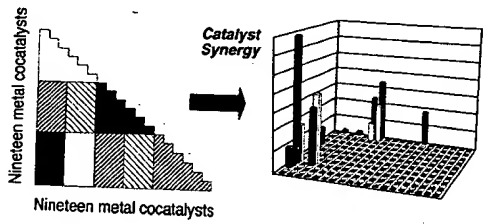

FIGURE 11. Full combinatorial design for two metals. The shading in the design indicates the blocking pattern used in the experiments.

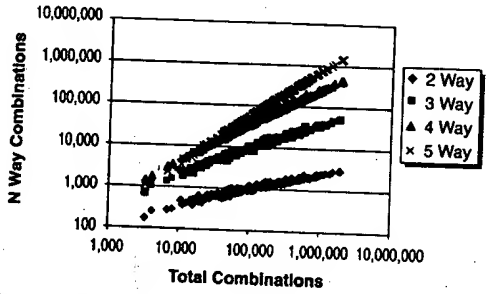

FIGURE 12. Total number of possible two-, three-, four-, and five-way combinations in a five-factor experiment with 2–20 levels per factor.

concentration level. If there had been an effect of relative concentrations, this, too, would be missed. Assessing the effect of concentrations on each metal combination would at least triple the number of tests.

Idealized Strategies. If we truly have no information on chemical possibilities, then the most effective strategy will be exhaustive enumeration of all n-way combinations of the factors. This leads to the first question: which of those combinations can be exhaustively studied in a practical experimental situation?

To investigate this question, we ran a series of simulations using Crystal Ball[38] software. The following example illustrates the results we have found:

Basic Parameters

- No chemical knowledge (descriptors) assumed
- Factors are independent (no nesting)
- Five factors, each with 2–20 levels
- Moderately high throughput experimentation; hundreds to thousands of runs feasible
- Experimental error negligible

Using this simulation, 250–1000 instances were calculated. The results are shown in Figure 12.

From Figure 12 we can draw some immediate conclusions. First, the total numbers of possible combinations in even as simple a system as this quickly rises into the hundreds of thousands or millions. It will not be possible to exhaustively test all the possibilities. Second, the number of two-way combinations stays in the low thousands and is therefore experimentally accessible with high-throughput technology. Third, while the four-way and five-way combinations rapidly increase to the hundreds of thousands and are relatively impractical, the three-way combinations, which remain in the low tens of thousands, appear accessible.

We gain further encouragement in this area when we note that the critical parameter is the number of experimental runs that must be performed, not the absolute number of combinations. In a five-factor combinatorial experiment, a single experimental run observes

- 10 two-way combinations
  (12, 13, 14, 15, 23, 24, 25, 34, 35, 45)
- 10 three-way combinations
  (123, 124, 125, 134, 135, 145, 234, 235, 245, 345)
- 5 four-way combinations
  (1234, 1235, 1245, 1345, 2345)

Therefore, the minimum required number of runs to observe all n-way combinations is much less than the total number of those combinations. In general, the theoretical minimum number of runs to observe all n-way combinations is the product of the number of levels of the n factors with the largest numbers of levels. Thus, the minimum runs

$$= l_{i,\max}\, l_{j,\max} \qquad \text{(two-way)}$$

$$= l_{i,\max}\, l_{j,\max}\, l_{k,\max} \qquad \text{(three-way)}$$

where $l_{i,\max}$ is the number of levels of the factor with the largest number of levels, $l_{j,\max}$ is the second largest, etc. The actual minimum may be slightly larger than the theoretical minimum in more irregular chemical spaces.

If we apply this calculation to the numbers of combinations found above, we discover that the number of runs (Figure 13) required to exhaustively study all possible two-way combinations is actually relatively small. Even three-way combinations become quite tractable. The figure also shows that in the three-way case there is a substantial advantage in working with experimental spaces with relatively equal number of levels.

Strategies for Observing Two-Way Combinations. Two-way combinations are relatively easy to observe, even in rather complex systems. The mathematical description for an array which exhaustively samples all two-way combinations of a set is an "orthogonal array of strength 2 and index 1".[39] The classical Latin square design is such an array which efficiently samples all two-way combinations in a symmetrical system, with all factors having the same number of levels (Figure 14).

Latin squares can be generalized to less symmetrical systems such as Youden squares, and orthogonal arrays of strength 2 are relatively easy to construct.

Strategies for Observing Three-Way Combinations. Three-way combinations are less easy to exhaustively observe. Although a "Latin cube" is possible for perfectly symmetrical systems, it is not generalizable. Orthogonal arrays of strength 3 and index 1, which would be required for these systems, are relatively rare and are quite difficult

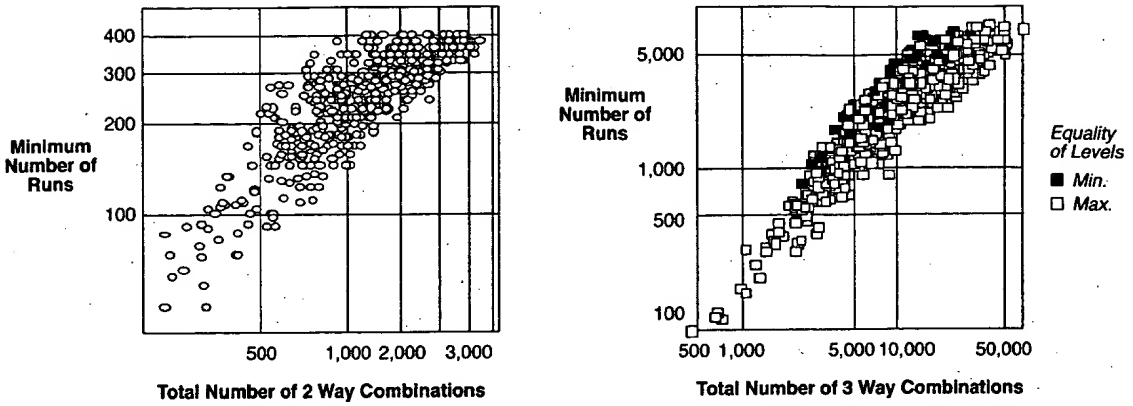

FIGURE 13. Minimum number of runs necessary to sample all two-way and three-way combinations in a five-factor design with 2–20 levels per factor.

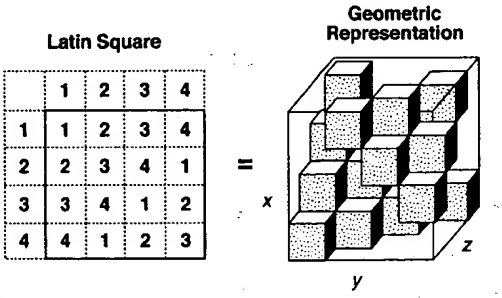

FIGURE 14. A Latin square observes all 64 two-way combinations of three factors with four levels each using 16 runs.

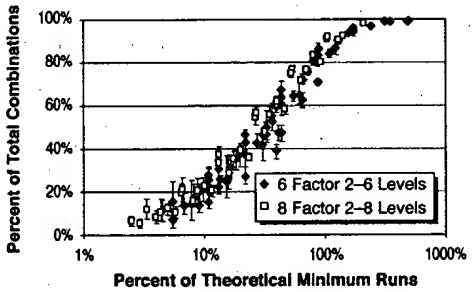

FIGURE 15. Three-way combinations observed with random runs. The results of each simulation are reported relative to the theoretical minimum runs for each combination of factors and levels. Each simulation was run at a given factor/level combination 10 times; the error bars show the range of the data.

to construct.[40] Therefore, algorithmic approaches are required. We have examined three strategies in this area:

- Random Runs
- Genetic Algorithms
- Computer-Generated Test Plans

1. Random Runs. The use of randomly chosen runs in a combinatorial study was investigated using simulation. The basic assumptions of the study were

- No chemical knowledge assumed
- Independent factors
- Six factors with 2–6 levels each or eight factors with 2–8 levels each

In each iteration of this simulation, a set of levels for each factor was randomly selected, and the list of all possible three-factor combinations was generated. Sets of 10, 20, 40, ... runs were then randomly generated and the resulting combinations checked off the list. The results of this simulation are given in Figure 15. It shows that random runs can be a relatively efficient method of sampling the three-way combinations in a fairly complex experiment. Approximately 80% of the combinations have been sampled by the time the theoretical minimum number of runs have been completed. Exhaustive sampling, however, is less successful; it requires about three times the theoretical minimum to sample 99% of the total combinations.

2. Genetic Algorithms. Genetic algorithms[41] (GAs) are a popular method of searching for optima in fields varying as widely as truck manufacturing and drug design. They have the advantage of being assumption free; they will work if there is any underlying structure to the experimental space—even if we cannot figure it out. The process of experimentation using genetic algorithms is straightforward:

- Selection of an experimental space consisting of compositional and process parameters which are combined to form a "genetic code" for producing the desired materials.
- Initialization of a first generation of materials. This is usually done by random selection, but it can be seeded with known "good" runs or constrained by prior knowledge.
- Preparation and testing of the materials from the first generation.
- Prioritizing the genetic codes from the first generation as "parents" for the next generation on the basis of the testing responses.
- Creation of the next generation from those parents by applying the evolutionary operators of crossover, qualitative mutation, and quantitative mutation. The critical design decisions in this methodology bear on the tradeoff between the rate of convergence on the best material vs the certainty of convergence. This is summarized in Table 4.[42] The principal disadvantage of GA

Table 4. Genetic Algorithm Design Factors

factors promoting high rate of convergence	factors promoting certainty of convergence
• small population • fitness-proportional selection (high reproduction rate of the best compositions) • quantitative mutation the prevalent evolutionary operator	• large population • all materials participate in the reproduction process, independently of performance • crossover and qualitative mutation the primary evolutionary operators

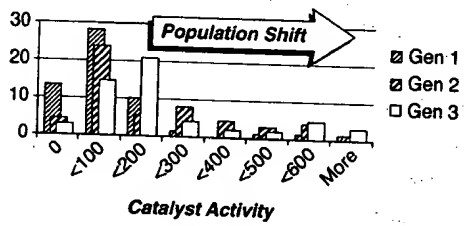

FIGURE 16. Three generations of a population of catalysts with 55 formulations per generation.

Table 5. Possible States of Automatic Telephone Software

call type	billing	access	status
local	caller	loop	success
long distance	collect	ISDN	busy
international	800	PBX	blocked

Table 6. Nine-Run Design That Tests All 54 Two-Way Combinations

	call type	billing	access	status
1	local	collect	PBX	busy
2	long distance	800	loop	busy
3	international	caller	ISDN	busy
4	local	800	ISDN	blocked
5	long distance	caller	PBX	blocked
6	international	collect	loop	blocked
7	local	caller	loop	success
8	long distance	collect	ISDN	success
9	international	800	PBX	success

strategies is the number of generations required for convergence. Most GA optimizations are run on computer models, so the cost and time required for running dozens to hundreds of generations are low. When a generational cycle requires a full process of running and analyzing an experiment, the cost may be too high. In a representative experiment run for three generations in our laboratory (Figure 16), there was a clear population shift toward higher activity, but the rate was too slow to be practical. In other (bio)chemical systems where GAs have been used, convergence or leveling of improvement has occurred in 10−20 generations.[43,44]

3. Computer-Generated Test Plans. With a computer-aided algorithm, it is possible to exhaustively enumerate all possible *n*-way combinations. This can be followed by selection of an appropriate subset of runs that will sample all *n*-way combinations. Fortunately, this problem has already been solved in another context—software test generation. Proper testing of software requires examination of combinations of inputs to test for untoward interactions. For example, an automatic telephone system might require examination of the possibilities shown in Table 5. There are 81 possible scenarios in this situation, which contain 54 possible two-way combinations. All of these combinations can be sampled in only nine experimental runs (Table 6).

A Web-based software service[45] has now been commercialized to generate such test plans. We have found it to be reasonably user-friendly and capable of accommodating complex experiments and constraints. For example, a catalyst system consisting of

- Primary catalyst: 4 possibilities
- Metal cocatalyst: 2 × 5 possibilities @ 2 concentrations
- Cocatalyst ligand: 6 possibilities
- Nonmetal cocatalyst: 3 possibilities @ 3 concentrations
- Process factors: 3 @ 2 levels

contained 6075 possible three-way combinations. The theoretical minimum number of runs to sample all three-way combinations is 150; the algorithm was able to find a 167-run plan that actually sampled them all.

The principal limitations of these test plans are the following:

- They are highly dependent on the significant interaction effects being synergistic rather than antagonistic. Even a modest poisoning effect can obliterate a large portion of the design.
- They require that the desired high-order interaction effect be relatively large, while the main effects and low-order interactions remain small. Otherwise, the desired observation will be drowned in the noise of the additive lower-order effects.
- The lack of redundancy requires that the quality of the experimental system be very high.

If these assumptions are met, a simple histogram or normal probability plot of the response data will identify the runs containing strongly positive interactions. If there is more than one such run, the active factors will be indicated by simple comparison. If there is only one, a followup design can be run with only two levels/factor to home in on the active factors. A resolution IV fractional factorial design (32 runs in the catalyst case above) will cover all the possible three-way combinations.[46]

Conclusion

These methods of high-throughput materials development are still in a rapid state of development, and experimental strategies appropriate to each methodology are also appearing rapidly. This Account does not delve into the full complexities of statistical analysis which may be required for some of these approaches; it is a very good idea to have an experienced statistician as a full member

of the team. Finally, the quality issues inherent in operation of an automated, high-throughput experimental system are substantial and will be discussed in a subsequent article.

The author thanks the GE Corporate R&D Combinatorial Chemistry program, led by Terry Leib, and the Applied Statistics Program, led by Gerald Hahn, for their contributions which helped make this review possible. Special thanks go to Charlie Hendrix (South Charleston, WV) for many fruitful discussions on the strategy of experimentation.

References

(1) Liu, D. R.; Schultz, P. G. Generating New Molecular Function: A Lesson From Nature. *Angew. Chem., Int. Ed.* **1999**, *38*, 36–54.
(2) Symyx Technologies, http://www.symyx.com (accessed Nov 2, 2000).
(3) Fisher, R. A. *The Design of Experiments*; Oliver and Boyd: Edinburg, 1935.
(4) Daniel, C. *Applications of Statistics to Industrial Experimentation*; John Wiley: New York, 1976.
(5) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters*; John Wiley: New York, 1978.
(6) Cohan, P. Achieving Strategic Change through Innovation. Presented at Combi 2000 (Knowledge Foundation), San Diego, CA, Jan 23–25, 2000.
(7) Brown, R. D.; Newsam, J. M. The importance of model behavior. *Chem. Ind. (London)* **1998**, 5 Oct, 785–788.
(8) Molecular Simulations Inc. Cerius2 environment, http://www.msi.com/materials/cerius2/index.html (accessed Nov 2, 2000).
(9) Caruana, C. M. Combinatorial Chemistry Promises Better Catalysts and Materials. *Chem. Eng. Prog.* **1998**, *94*, 11–14.
(10) Reddington, E.; Sapienza, A.; Gurau, B.; Viswanathan, R.; Sarangapani, S.; Smotkin, E. S.; Mallouk, T. E. Combinatorial Electrochemistry: A Highly Parallel, Optical Screening Method for Discovery of Better Electrocatalysts. *Science* **1998**, *280*, 1735–1737.
(11) Gordon, E. M.; Gallop, M. A.; Patel, D. V. Strategy and Tactics in Combinatorial Organic Synthesis. Applications to Drug Discovery. *Acc. Chem. Res.* **1996**, *29*, 144–154.
(12) Young, S. S.; Sheffield, C. F.; Farmen, M. Optimum Utilization of a Compound Collection or Chemical Library for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 892–899.
(13) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Walden, M. Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Diversity* **1996**, *2*, 64–74.
(14) Clark, R. D.; Cramer, R. D. Taming the Combinatorial Centipede. *CHEMTECH* **1997**, *1997*, 24–31.
(15) Young, S. S.; Farmen, M.; Rusinko, A., III. Random Versus Rational Which is Better for General Compound Screening? http://www.netsci.org/Science/Screening/feature09.html (accessed Nov 2, 2000).
(16) Martin, E. J.; Critchlow, R. E. Beyond Mere Diversity: Tailoring Combinatorial Libraries for Drug Discovery. *J. Comb. Chem.* **1999**, *1*, 32–45.
(17) Reynolds, C. H.; Druker, R.; Pfahler, L. Lead Discovery Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 305–312.
(18) Warr, W. A. Combinatorial Chemistry and Molecular Diversity. An Overview. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 134–140.
(19) Agrafiotis, D. K.; Myslik, J. C.; Salemme, F. R. Advances in diversity profiling and combinatorial series design. *Mol. Diversity* **1999**, *4*, 1–22.
(20) Burgess, K.; Moye-Sherman, D. A.; Porte, M. In *Molecular Diversity and Combinatorial Chemistry*; Chaiken, I. M., Chanda, K. D., Eds.; American Chemical Society: Washington, DC, 1996; pp 128–136.
(21) Hanak, J. J. The "Multi-Sample Concept" in Materials Research: Synthesis, Compositional Analysis and Testing of Entire Multicomponent Systems. *J. Mater. Sci* **1970**, *5*, 964–971.
(22) Van Dover, R. B.; Schneemeyer, L. F.; Fleming, R. M. Discovery of a useful thin-film dielectric using a composition-spread approach. *Nature (London)* **1998**, *392*, 162–164.
(23) Senkan, S.; Krantz, K.; Oztrurk, S.; Zengin, V.; Onal, I. High-Throughput Testing of Heterogeneous Catalyst Libraries Using Array Microreactors and Mass Spectrometry. *Angew. Chem., Int. Ed.* **1999**, *38*, 2794–2799.
(24) Cong, P.; Doolen, R. D.; Fan, Q.; Giaquinta, D. M.; Guan, S.; McFarland, E. W.; Poojary, D. M.; Self, K.; Turner, H. W.; Weinberg, H. High-Throughput Synthesis and Screening of Combinatorial Heterogeneous Catalyst Libraries. *Angew. Chem., Int. Ed.* **1999**, *38*, 484–488.
(25) Sun, T. X. Combinatorial Search for Advanced Luminescent Materials. *Biotechnol. Bioeng.* **1999**, *61*, 193–201.
(26) Wang, J.; Yoo, Y.; Gao, C.; Takeuchi, I. S., X.; Chang, H.; Xiang, X.-D.; Schultz, P. G. Identification of a Blue Photoluminescent Composite Material from a Combinatorial Library. *Science* **1998**, *279*, 1712–1714.
(27) Xiang, X.-D.; Sun, X.; Briceno, G.; Lou, Y.; Wang, K.-A.; Chang, H.; Wallace-Freedman, W. G.; Chen, S.-W.; Schultz, P. G. A Combinatorial Approach to Materials Discovery. *Science* **1995**, *268*, 1738–1740.
(28) Schmitz, C.; Posch, P.; Thelakkat, M.; Schmidt, H.-W. Efficient screening of electron transport material in multilayer organic light emitting diodes by combinatorial methods. *Phys. Chem. Chem. Phys.* **1999**, *1*, 1777–1781.
(29) Cornell, J. A. *Experiments With Mixtures: Designs, Models, and the Analysis of Mixture Data*, 2nd ed.; John Wiley and Sons: New York, 1990.
(30) Montgomery, D. C. In *Design and Analysis of Experiments*; John Wiley & Sons: New York, 1984; pp 357ff.
(31) Milliken, G. A.; Johnson, D. E. *Analysis of Messy Data*; Van Nostrand Reinhold: New York, 1984.
(32) Hendrix, C. *NESTED97*; South Charleston, WV, 1997.
(33) Bisgaard, S. The Design and Analysis of 2∧(k−p) x 2∧(q−r) Split Plot Experiments. *J. Qual. Tech.* **2000**, *32*, 39–56.
(34) Crabtree, R. H. Combinatorial and rapid screening approaches to homogeneous catalyst discovery and optimization. *Chem Commun.* **1999**, 1611–1616.
(35) Shimuzu, K. D.; Snapper, M. L.; Hoveyda, A. H. High-Throughput Strategies for the Discovery of Catalysts. *Chem. Eur. J.* **1998**, *4*, 1885–1889.
(36) Kreuger, C. A.; Snapper, M. L.; Hoveyda, A. H. In *Abstracts of Papers*, 216th National Meeting of the American Chemical Society, Boston, MA, Fall 1998; American Chemical Society: Washington, DC, 1998; p 232.
(37) Cawse, J. N.; Doganaksoy, N.; Hansen, C.; Mattheyses, R.; Pisupati, C.; Repoff, T.; Stanard, C.; Tucker, W. Presented at the ASA Quality & Productivity Research Conference, Schenectady, NY, 1998.
(38) Decisioneering Advanced Analytic Tools, http://www.decisioneering.com/ (acessed Nov 2, 2000).
(39) Colbourn, C. J.; Dinitz, J. H. In *The CRC Handbook of Combinatorial Designs*; Colbourn, C. J., Dinitz, J. H., Eds.; CRC Press: New York, 1996; pp 97ff.
(40) Kreher, D. R., Michigan Technological University, Houghton, MI, personal communication, 1998.
(41) Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: New York, 1989.
(42) Baerns, M. Presented at Combi 2000 (Knowledge Foundation), San Diego, CA, 2000.
(43) Weber, L.; Wallbaum, S.; Broger, C.; Gubernator, K. Optimization of the Biological Activity of Combinatorial Compound Libraries by a Genetic Algorithm. *Angew. Chem., Int Ed.* **1995**, *34*, 2280–2282.
(44) Sheridan, R. P.; Kearsley, S. K. Using a Genetic Algorithm to Suggest Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1994**, *35*, 310–320.
(45) AETG Web, http://aetgweb.argreenhouse.com/ (accessed Nov 2, 2000).
(46) Street, D. J. In *The CRD Handbook of Combinatorial Designs*; Colbourn, C. J., Dinitz, J. H., Eds.; CRD Press: New York, 1996; p 339.

AR000117S